Journal of proteome research

pubs.acs.org/jpr

Article

Protein−Protein Interaction Networks Derived from Classical and Machine Learning-Based Natural Language Processing Tools

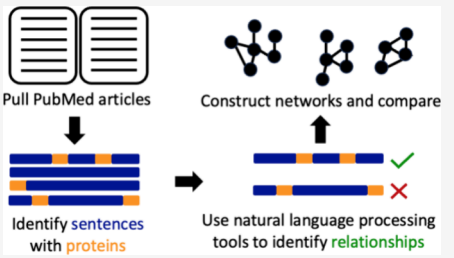
David J. Degnan, Clayton W. Strauch, Moses Y. Obiri, Erik D. VonKaenel, Grace S. Kim, James D. Kershaw, David L. Novelli, Karl TL Pazdernik, and Lisa M. Bramer*

Cite This: *J. Proteome Res.* 2024, 23, 5395−5404

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The study of protein−protein interactions (PPIs) provides insight into various biological mechanisms, including the binding of antibodies to antigens, enzymes to inhibitors or promoters, and receptors to ligands. Recent studies of PPIs have led to significant biological breakthroughs. For example, the study of PPIs involved in the human:SARS-CoV-2 viral infection mechanism aided in the development of SARS-CoV-2 vaccines. Though several databases exist for the manual curation of PPI networks, text mining methods have been routinely demonstrated as useful alternatives for newly studied or understudied species, where databases are incomplete. Here, the relationship extraction performance of several open-source classical text processing, machine learning (ML)-based natural language processing (NLP), and large language model (LLM)-based NLP tools was compared. Overall, our results indicated that networks derived from classical methods tend to have high true positive rates at the expense of having overconnected networks, ML-based NLP methods have lower true positive rates but networks with the closest structures to the target network, and LLM-based NLP methods tend to exist between the two other approaches, with variable performances. The selection of a specific NLP approach should be tied to the needs of a study and text availability, as models varied in performance due to the amount of text provided.



KEYWORDS: *natural language processing, relationship extraction, biological text mining, machine learning, large language models, BERT, GPT, LLM*

INTRODUCTION

Proteins interact with other proteins to fulfill highly complex biological functions, such as catalyzing chemical reactions, regulating cellular functions, controlling gene expression, and forming structural elements.[1] Modern studies of protein−protein interactions (PPIs) have elucidated previously poorly understood relationships, like the SARS-CoV-2 infection mechanism of coronavirus spike proteins and human ACE2, which have led to important developments like vaccines.[2] Besides offering insight into protein targets during vaccine development, PPI networks have aided in the discovery of novel protein functions, conserved molecular interaction patterns across species or disease states, and causal modeling where the effect of removing or modifying a particular protein on a biological system is studied.[3] Therefore, understanding PPIs is crucial to obtaining biological discoveries. Many widely used databases exist for storing experimentally derived and manually curated PPIs for multiple organisms, including STRING,[4] IntAct,[5] UniProt,[6] and several others. As the number of studied biological species and systems continues to increase, the expectation that a database should house all PPIs for every biological species, strain, and disease state has become unreasonable. Thus, text mining and natural language processing (NLP) approaches have become commonly used to extract PPIs from the literature, from text-mining "classical" approaches to simple neural nets to deep learning approaches (including ones that detect terms with named entity recognition) and large language models (LLMs).[7−19]

Early text mining methods, ones referred to as "classical" in this publication, rely on the co-occurrence of two proteins in a sentence or abstract.[7,8] The Chilibot[7] method identifies relationships as those with a specific relational term (e.g., binds, inhibits, promotes) within the same sentence as the two proteins of interest. Another classical method, pubmed.miner,[8] scores potential relationships using a cosine correlation of the counts of terms across contexts, like abstracts. Alongside the classical methods, several simple neural network models emerged that used word embeddings and oftentimes combinations of pattern analysis algorithms called kernels to identify related terms, with promising performances.[9−12] By the late 2010s, more complex machine learning (ML)-based methods with several layers emerged, often referred to as "deep learning" models. Some methods, like REACH[13] and TRIPS,[14]

Received: June 23, 2024
Revised: October 15, 2024
Accepted: November 1, 2024
Published: November 11, 2024

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404

Table 1. Overview of Selected Open-Source Natural Language Processing Tools and Their Inputs, Year Published, Link and Their Type[a]

algorithm	algorithm type	inputs	year published	algorithm link
Sentence Co-Occurrence[7]	classical	sentences, terms	2005	github.com/pnnl-predictive-phenomics/biological_relationship_extraction
Relational Term[7]	classical	sentences, terms	2005	github.com/pnnl-predictive-phenomics/biological_relationship_extraction
Fixed Term[7]	classical	sentences, terms	2005	github.com/pnnl-predictive-phenomics/biological_relationship_extraction
pubmed.mineR and cosine[8]	classical	sentences, terms, score threshold	2015	https://cran.r-project.org/web/packages/pubmed.mineR/index.html
TRIPS[14]	ML	sentences only	2015	https://github.com/sorgerlab/indra
REACH[13]	ML	sentences only	2018	https://github.com/sorgerlab/indra
PubMedBERT[16]	ML	sentences, terms	2019	https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext
BioBERT[17]	ML	sentences, terms	2020	https://huggingface.co/dmis-lab/biobert-base-cased-v1.2
BioGPT[18]	LLM	sentences, terms	2022	https://huggingface.co/microsoft/BioGPT-Large-PubMedQA
SOLAR[19]	LLM	sentences, terms	2023	https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0
Gemini Pro 1.5[20]	LLM	sentences, terms	2024	https://ai.google.dev/

[a]Where "ML" stands for machine learning, and "LLM" stands for large language model, which is a type of machine learning algorithm.

paired relationship extraction (i.e., does the relationship exist or not) with biological named entity recognition, meaning that users only need to supply a context, and all possible terms and relationships are then extracted from that context. Other methods, like PubMedBERT[16] and BioBERT,[17] were versions of the BERT[20] model that were trained on publications from biomedical databases like PubMed and require further training to be tuned for relationship extraction tasks. Typically, these models require a sentence and two terms (in this context, the name of the two proteins that could have a relationship) and will return a probability that the first term is related to the second.

Since 2023, several deep-learning-based question answering (QA) LLMs have emerged, including BioGPT,[18] SOLAR,[19] and Google's Gemini.[19] Though these models do not require additional training, they behave differently than other ML models in terms of both input (a question with context, which is text pulled from a publication) and output (a response that needs to be collapsed into a "the relationship exists" or "the relationship does not exist") and thus must be handled appropriately.[21]

Here, a selection of tools from the vast landscape of NLP tools were compared in three studies. Each tool's ability to properly identify PPIs or build PPI networks was tested with (1) a set of hand-verified relationships,[22] (2) an experimentally and expert-reviewed network pulled from UniProt[6] for *C. elegans*, and (3) a PubMed query for papers meant to identify *E. coli* proteins involved in metabolism, with no target network to simulate an example where the truth is not known. True positive (the algorithm correctly predicted an explicitly mentioned relationship) and true negative (the algorithm correctly predicted that a relationship was not explicitly mentioned) rates were calculated for each model, and performance metrics were calculated (Table S1). Network-specific metrics were also calculated, including connectedness,[23,24] the clustering coefficient,[25−28] the number of components,[29] the average component size,[29,30] the average neighborhood size,[30] and the graph edit distance.[31,32] Top performing algorithms per category (classical, non-LLM ML, or LLM) were then reported. All code for this study is available at: github.com/pnnl-predictive-phenomics/biological_relationship_extraction.

METHODS

Description of True Positives, True Negatives, False Positives, and False Negatives

To consider two proteins related, the provided context (e.g., a sentence) must have an explicit relationship expressed between each term (Figure S1). Any implied or not mentioned relationship was considered a nonrelationship. Thus, true positives are relationships that are explicitly mentioned in a context and labeled as such relationships by a tool. True negatives are nonrelationships that are not explicitly mentioned in a context and labeled as nonrelationships by a tool (Figure S1). False positives are cases when relationships are incorrectly identified as relationships, and conversely, false negatives are cases when relationships are incorrectly identified as not relationships (Figure S1). In the example "ACES binds COLQ and ENO1″, there is an explicit relationship between ACES:COLQ and ACES:ENO1 (the positive cases), and there is not an explicit relationship mentioned between COLQ:ENO1 (the negative cases). Therefore, if ACES:COLQ was labeled by a tool as a not relationship, it would be a false negative, and if COLQ:ENO1 was labeled by a tool as a relationship, it would be a false positive.

NLP Tool Selection and Description

Selected algorithms were required to be open-source, accessible in Python or R, and have the capability to identify biological relationships as either supported by the provided text (context) or not. We selected tools that represented popular approaches to biological text mining: earlier text-matching methods referred to as classical methods, ML-based methods (non-LLM), and LLMs (Table 1). The previously mentioned simple neural network approaches[9−12] did not meet these requirements and were not included in this study.

Three classical methods were derived from Chilibot.[7] "Sentence Co-Occurrence" flagged two terms as related if they co-occur in the same sentence. "Relational Term" expanded on Sentence Co-Occurrence by requiring a relational term (e.g., binds, inhibits, etc.) between the two candidate proteins. A full list of selected relational terms was designed and agreed on by a cohort of 10 experts and can be found here: https://github.com/pnnl-predictive-phenomics/biological_relationship_extraction/blob/main/algorithms/co_occurrence.R#L23-L42. "Fixed Term" further expanded the

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404

requirements of the Relational Term by requiring the relational term to be positioned between the two candidate proteins. The last classical method, pubmed.mineR,[8] utilized a cosine correlation score in a count matrix of terms where each column is a term, each row is a context (a sentence in this study), and the values are the counts of each term. Here, abstracts or full papers were split into sentences, count matrices were made, and a cosine correlation was calculated between the counts. Receiver operating characteristic (ROC)[33] curves were drawn for cosine correlations to determine an optimum score cutoff for distinguishing whether two proteins were related or not using a 1:1 cost where true positives and true negatives are weighted equally. The formula for calculating cost was calculated by eq 1, where R was the number of relationships, T was the total number of queries, TN was the number of true negatives, FP was the number of false positives, TP was the number of true positives, and FN was the number of false negatives.

$$1{:}\,1\ \text{cost} = \left(1 - \frac{R}{T}\right)\frac{\text{TN}}{\text{FP} + \text{TN}} + \frac{R}{T}\left(1 - \frac{\text{TP}}{\text{TP} + \text{FN}}\right) \quad (1)$$

Eq 1 shows the formula for calculating a 1:1 cost to determine an optimum cosine correlation score threshold for pubmed.mineR.

Four non-LLM ML models were selected for this study. Two were context-only algorithms (TRIPS[14,15] and REACH[13,15]) where biological terms are detected and extracted by the tools. Both were made accessible in Python by the INDRA[15] project. The other two algorithms, PubMedBERT[16,34] and BioBERT,[17,34] required both context and terms. BERT models were utilized from Hugging Face (Table 1), and were trained with the data provided by Su and Vijay.[34] The Su training data set contained 5818 sentences, where 4818 included two terms that were not related and 1000 included terms that were related, encoded as a 0 or 1.[34] The code was adapted from Lee et al.,[22] where the authors characterized biological relationships as belonging to one of eight categories. In this context, the BERT models were trained to classify context as two categories—related or not, based on the example data provided by Su and Vijay. On the training data set from Su and Vijay,[34] BioBERT had a total F1 score of 0.929, and PubMedBERT had a total F1 score of 0.931.

Three LLM models were tested: BioGPT,[18] SOLAR,[19] and Google's Gemini Pro[19] version 1.5. A specific instance of BioGPT that had been trained on question-answering tasks from PubMed data, BioGPT-Large-PubMedQA, was used. SOLAR's 10.7 billion parameter model was tested using API calls to a local instance of text-generation-webui: https://github.com/oobabooga/text-generation-webui. Gemini was accessed via the google.generativeai[19] python library. It should be noted that in the free tier of Gemini, only 50 queries can be made to Google's API service a day, at the time of this writing. To keep contexts consistent across all algorithms for comparisons, all LLMs were provided with the same sentences as the other models. The format for the queries used was: "Context: [Context]. Question: Based solely on the provided context, does [Protein 1] interact with [Protein 2]? Answer with a "yes" or "no"." LLMs were not further trained, as training them is not a requirement for using them for biological relationship extraction, as opposed to the BERT models, where additional training is required to use them for biological relationship extraction. All LLM models were run with a default temperature of 1.0.

Study 1: Relationship Data Sets

Two data sets were used to test the relationship extraction capabilities of each algorithm—one manually curated data set of gene products interacting with other gene products (GP-GP) and a published data set, BioRED.[35] The GP-GP data set comprised 25 abstracts where every possible gene product's relationship was paired with every other possible gene product and manually annotated as a relationship or not. This resulted in 519 annotations, 199 relationships, and 320 nonrelationships. Hand annotations were verified and reached a singular consensus by three experts.

BioRED[35] relationships were filtered down to only PPIs, and negatives were generated by taking any relationships not explicitly annotated by BioRED, as long as both terms occurred within the same sentence. This resulted in 291 relationships and 52 nonrelationships across 77 abstracts.

Study 2: Pulling Publications, Extracting Sentences, and Building the Target Network for *C. elegans*

In the case of the *C. elegans* data set, known PPIs were exported from UniProt[6] with their respective PubMed paper IDs (PMIDs) using the UniProtR[36] R package. Briefly, these binary interactions are from the IntAct[5] database and have been filtered to remove those only supported by one experiment, ones that only exist in large complexes, and ones only inferred in literature but not experimentally demonstrated. Thus, all interactions designated as true in this study are validated by more than one experiment.

Literature was pulled from PubMed using a custom-built Python[37] pipeline that used PMIDs as input, found here: github.com/pnnl-predictive-phenomics/biological_relationship_extraction/. Full-text publications were pulled as either "clean text" or as PDFs. Clean text is human readable and removes extra information that may not be relevant that is found in PDFs, including tables and hexacodes for color. Clean text, whenever available, was parsed directly from XML, which is stored in the PubMed archives.[38] These archives were accessed and parsed with the BeautifulSoup4[39] and requests[37] libraries via the PubMed Central File Transfer Protocol service. PDFs were pulled with the MetaPub (https://github.com/metapub/metapub) library, and text was extracted from the files with the pypdf (https://github.com/py-pdf/pypdf) library. Full text was not always available for a variety of reasons, including closed access publications and authentication checks that do not permit automated text extraction. In cases where no full text was available, titles and abstracts for publications were pulled directly from the PubMed Web site using BeautifulSoup.[39] Both titles and abstracts are publicly available with only a handful of exceptions, such as paper correction memos that are listed with PMIDs but are not publications. Across the 1446 papers identified as containing PPIs for *C. elegans* by UniProt, 314 clean text, 475 PDFs, and 654 titles and abstracts were pulled. Only three publications did not have either full text or abstracts and were not used in the study.

To map proteins in this text to their UniProt protein IDs, a synonym dictionary for each protein was created, where the first column contained the UniProt protein ID and the second column contained a corresponding common name term—called a "synonym". These synonyms were pulled from UniProt, and cases where multiple IDs mapped to a synonym

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404

Figure 1. Performance of algorithms based on balanced accuracy (BA), true positive rate (TPR), and true negative rate (TNR) on the GP-GP (left) and BioRED (right) data sets. Algorithms are ordered by the GP-GP BA.

or cases where synonyms had less than 3 characters were filtered out. Next, UniProt IDs were identified in PubMed papers by their synonyms (Figure S2). Any sentence in the papers that had at least two UniProt protein IDs was returned and tested by each NLP model. The "target" network was derived by filtering the UniProt *C. elegans* PPI network down to all PPIs that were identified in these sentences (Figure S2). All NLP models were compared to this target network to determine the model performance.

Comparison Studies of Clean Text versus PDFs and Full Text versus Abstract

To determine whether PDFs or "clean text" (a cleaner version of PDFs without tables, hexacodes, etc.) perform better as full text for NLP tasks, a cohort of 305 *C. elegans* papers were pulled as both PDFs and clean text. The cohort size was based on the full text availability. Both cohorts were run through a representative model from the classical, ML, and LLM model types, which were Sentence Co-Occurrence (classical), PubMedBERT (ML), and SOLAR (LLM). Results were then compared to the *C. elegans* target network, as described previously.

Next, the same algorithms were used to compare networks derived from full text articles ("clean text" and PDF) and paper titles and abstracts (n = 785). For comparison purposes, the same *C. elegans* target network is used.

Study 3: Pulling Publications and Extracting Sentences for *E. coli*

In the case of *E. coli*, PubMed was queried with the search term [(*E. coli* proteomics) AND (*E. coli* metabolism) AND (("2000/01/01"[Date - Publication]: "2024/02/23"[Date - Publication]))], which means that all papers with the terms "*E. coli* proteomics" and "*E. coli* metabolism" published from January 1st, 2000 to February 23rd, 2024 were pulled. This resulted in 3846 papers: 1160 full text (1038 as "clean text" and 122 as PDFs) and the remaining 2686 were pulled as titles and abstracts, as described previously. UniProt was used to map protein IDs to common names, and sentences with at least two UniProt protein IDs in them were extracted and passed to NLP tools (Figure S2).

Standardizing Predictions across NLP Tools

All predicted PPIs were converted to their UniProt IDs using the synonym table and collapsed to unique instances of each PPI, since the interaction "Protein A:Protein B" was considered the same as the interaction of "Protein B:Protein A" in this study. If, in any sentence, a PPI was identified by an NLP tool, then the PPI was returned as "identified" by the tool. This means that if there are multiple sentences with two terms, a PPI will be identified by an NLP tool if it was identified in any one of the sentences. In other words, there needs to be only one supported context for a PPI to be considered predicted by an NLP tool. This method was applied to the outputs of all NLP tools for consistent comparisons.

Metrics

True negative cases were only considered in study 1 and not studies 2 and 3, as most PPI networks are incomplete based on data availability in manually curated databases.[3] In study 1, true negative rates (TNRs) were then calculated as the number of true negatives divided by the number of false positives added to the number of true negatives. True positive rates (TPRs) were calculated as the number of true positives divided by the number of false negatives added to the number of true positives. In study 1, balanced accuracy (BA) was used to rank algorithm performances, which are the average of the true positive and true negative rates. BA was selected over the F1-score, as BA prescribes an equal weight to true positives and true negatives whereas the F1-score does not weight true negatives. In this study, there was no convincing literature to weight true positives more than true negative cases.

Other metrics were used to compare network structures in studies 2 and 3. Connectedness is part of Krackhardt's four

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404

Table 2. Balanced Accuracy (BA), True Positive Rates (TPR), and True Negative Rates (TNR) for the GP-GP (Gene Product-Gene Product) and BioRED[35] Datasets, Sorted by the GP-GP Balanced Accuracy

algorithm	model type	GP-GP BA	GP-GP TPR	GP-GP TNR	BioRED BA	BioRED TPR	BioRED TNR
Fixed Term	classical	0.8105	0.6834	0.9375	0.7636	0.6426	0.8846
Relational Term	classical	0.7319	0.8794	0.5844	0.6778	0.8557	0.5000
Gemini Pro 1.5	LLM	0.7030	0.4372	0.9688	0.7993	0.6564	0.9423
Co-Occurrence	classical	0.6741	0.9045	0.4438	0.6731	1.0000	0.3462
pubmed.mineR + Cosine 0.03	classical	0.6741	0.9045	0.4438	0.6731	1.0000	0.3462
SOLAR	LLM	0.6697	0.6332	0.7063	0.5210	0.5514	0.4906
PubMedBERT	ML	0.5537	0.1106	0.9969	0.5515	0.1031	1.0000
BioBERT	ML	0.5452	0.0905	1.0000	0.5567	0.1134	1.0000
REACH	ML	0.5371	0.0804	0.9938	0.6151	0.2302	1.0000
TRIPS	ML	0.5000	0.0000	1.0000	0.5000	0.0000	1.0000
Pubmed.mineR + Cosine 0.5	classical	0.4970	0.4221	0.5719	0.4978	0.4570	0.5385
BioGPT	LLM	0.4787	0.2261	0.7313	0.4764	0.2027	0.7500

measures for summarizing hierarchical structures,[23] and is the number of connected nodes divided by the total number of possible edges in the network.[23,24] Higher values indicate more connected networks. The average global clustering coefficient ranges between 0 and 1, where higher values represent more connected triplets of nodes in the network.[25−28] This value is calculated as two times the number of triangles attached to a node divided by the degree of the node multiplied by the degree of the node minus one.[26] The number of components is the number of distinct sections of interconnected nodes (disjoint sets),[29] and the average component size is the average number of nodes per component.[29] The average neighborhood size[30] is the average number of edges per node. Finally, the graph edit distance is a metric of similarity between two graphs, where lower values indicate a more similar network.[31,32]

RESULTS AND DISCUSSION

Cosine Similarity Score Thresholds for pubmed.mineR

To determine an optimum cosine correlation score, a 1:1 cost was used, where true positives and true negatives were weighted equally. For the GP-GP data set, BioRED data set, and the *C. elegans* study where a target network is known, the optimum correlation score was shrunk toward the smallest possible score (Figure S3). This indicates that one co-occurrence of two terms in an article was as predictive of a PPI as multiple within the same article. Interestingly, this smallest score performs the same as sentence cooccurrence (Figure 1). In study 2, a cosine correlation threshold of 0.5 was used with pubmed.mineR to distinguish its performance from sentence co-occurrence.

Study 1: Relationship Extraction on Cohort of Abstracts

On the GP-GP data set, classical methods tended to perform better than ML and LLM approaches based on BA, as the median rank for the classical approaches was 4 and for the ML/LLM approaches was 8. The top two performing algorithms were fixed term (0.8105) and related term (0.7319), both classical methods, and the third top performing algorithm was the ML/LLM approach Gemini (0.7030) (Figure 1). Overall, classical methods tended to have higher TPRs at the expense of lower TNRs, and ML/LLM methods had lower TPRs and higher TNRs (Table 2). Only three algorithms performed near or below random, which contained one classical method (pubmed.mineR with a cosine threshold of 0.5) and two ML/LLM methods: TRIPS and BioGPT.

There were some differences in performance between GP-GP and BioRED, with the largest differences occurring in the ML/LLM approaches with REACH, Gemini, and SOLAR. REACH and had higher BAs on the BioRED data set, which may be due to the presence of more common protein names in BioRED that REACH can identify with its named entity recognition capabilities, as opposed to GP-GP which mixed commonly known and less common proteins and gene products. The difference in the performance of Gemini and SOLAR may be explained by the inherent randomness of LLMs (e.g., their creativity), which makes characterizing their behavior difficult.

Examples of false negatives and false positives across tool types were explored. In terms of false positives, the "classical methods" that were derived from Chilibot[7] (Co-Occurrence, Relational Term, and Fixed Term) failed whenever a relational term was between the two terms even though there was no relationship. For example, "the expression of these proteins was also analyzed with ··· [protein 1] and proliferation-associated [protein 2]"[40] had the term "associated" between protein 1 and 2, indicating that protein 2 had proliferation properties and not that protein 1 interacted with protein 2. False positives for the count-based classical method pubmed.mineR were simply cases where two terms were mentioned several times together in an abstract, even though they were not related. For example, in the abstract from Urade et al., "CNX" and "CRT" co-occurred several times throughout the abstract, though the publication was about how each bound to another protein "ER-60" and not to each other.[41] In terms of the ML models, false positive cases were rare, as REACH had two false positives, PubMedBERT had one, and BioBERT had none. These false positive cases were instances with complex language patterns. For example, "···ANG II is linked to the activation of [protein 1] receptors and involves PKC activation and upregulation of renal [protein 2] but not of HO2 protein expression",[42] which contains multiple terms to indicate a relationship (e.g., activation, upregulation, expression) though an explicit relationship between protein 1 and 2 was not specified. False positives in LLM methods appeared to be cases in which a relationship may be implied but not explicitly stated. For example, "··· production of anti-inflammatory cytokines [protein 1]··· was increased and production of inflammatory chemokines [protein 2]··· was reduced",[43] which does not explicitly state a relationship between protein 1 and protein 2.

In terms of false negatives, the Chilibot[7] methods were either cases where terms did not co-occur in a sentence or no

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404

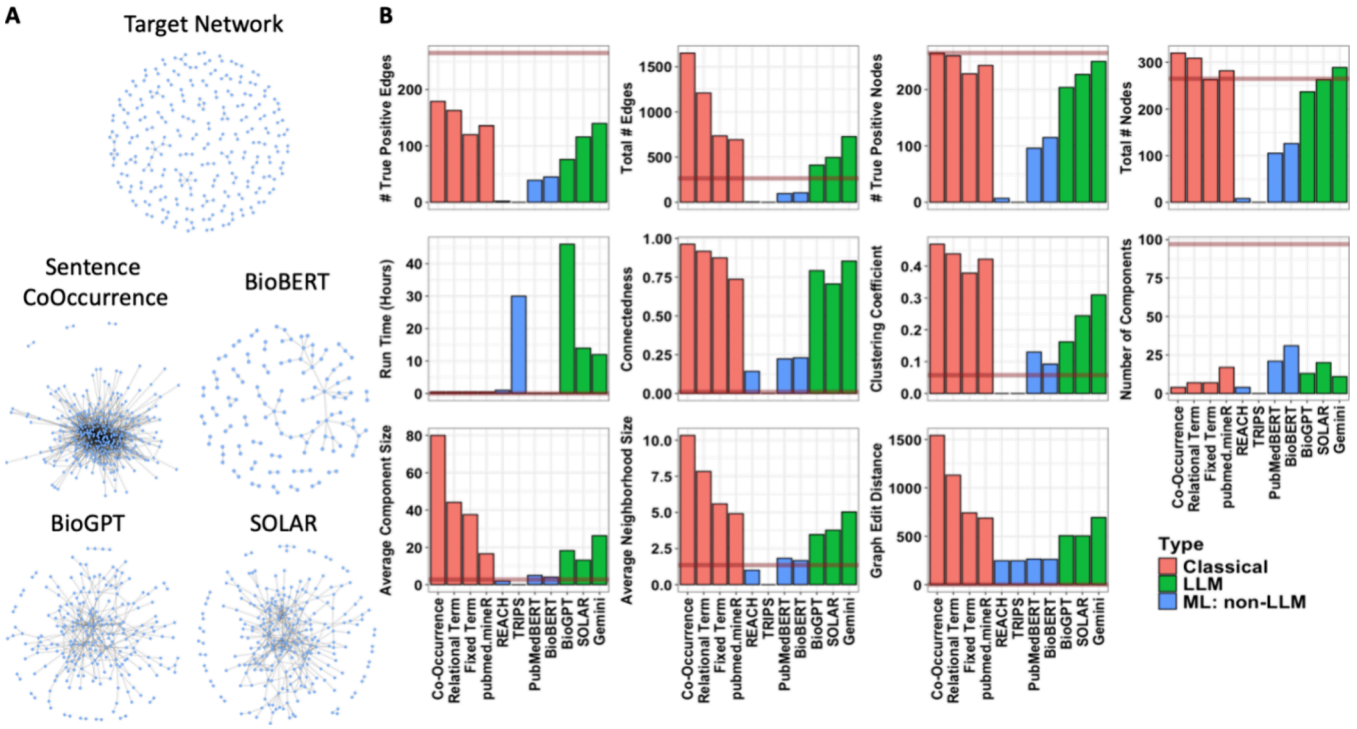


Figure 2. (A) Truth (top) and predicted (middle and bottom) *C. elegans* protein−protein interaction networks for a selection of algorithms. (B) Metrics evaluating each predicted network's quality and build time, with a red horizontal line to indicate the target value for each metric.

relational term was between related terms (for the Relational Term and Fixed Term methods). For example, "[protein 1] expression was correlated with tumor grade as well as with [protein 2]"[40] does not contain a relational verb and was thus incorrectly predicted as a negative case. For pubmed.mineR, terms that did not co-occur several times throughout an abstract were considered negatives. False negative cases for ML/LLM methods were difficult to diagnose, though examples that all methods of these types (PubMedBERT, BioBERT, REACH, TRIPS, SOLAR, BioGPT, and Gemini) missed tended to have contextual information that was brief or in a list with other proteins, such as "[protein 1] signaling regulates [protein 2] expression",[44] "[protein 1] significantly induced expression of ···TNFalpha, [protein 2], and IL-1beta",[45] and "[protein 1] co-stimuli enhanced IL-2, IFN-gamma, or [protein 2]".[45]

Comparison of the Clean Text and PDF Extraction Methods

To determine an optimum format for full text, 305 publications were compared with their PDFs and "clean text" versions, as previously described. The overall networks were quite similar (Figure S4a), with slight differences in performance. The TPRs were slightly higher (<0.01) for PDFs using the Sentence Co-Occurrence and PubMedBERT methods (Figure S4b). Interestingly, SOLAR had an increase in its TPR (increase of 0.05) when using clean text over PDFs. Connectedness scores, which measure how connected the nodes are, were closer to the target value for clean text in the Sentence Co-Occurrence (increase of 0.11) and PubMedBERT (increase of 0.13), but further from the target value for the SOLAR (decrease of 0.17). The results of this comparison did not provide a compelling argument for the usage of one full text file type over another; thus, clean text was used over PDF whenever possible, as clean text is more human readable than PDF.

Comparison of Full Text and Titles and Abstracts

To understand the effect of context size on tool performance, titles, abstracts, and full text articles were compared for 785 papers. All tested tools (Sentence Co-Occurrence, PubMedBERT, and SOLAR) demonstrated large improvements in TPRs from the usage of full text, with an increase of 0.7, 0.16, and 0.49, for each algorithm, respectively (Figure S5). This increase in TPR was at the detriment of each resultant network's connectedness score, as each method had increased scores that were off the target value of 0.0096. Both Co-Occurrence and SOLAR had large connectedness scores that were far off target for the "full text" case, with scores of 0.96 and 0.70, respectively. Noticeably, PubMedBERT had the smallest increase in connectedness, from 0.09 to 0.18, and the closest connectedness score to the target value. Given these results, in this study, full text was provided to algorithms wherever possible to increase TPRs with the understanding that some algorithms will have overconnected networks.

Study 2: Tool Performances on the *C. elegans* Paper Cohort

Predicted networks (Figure S6) from each tool were compared for their similarity to the target network for *C. elegans* (Figure 2a). No model type closely aligned with the target network for every metric (Figure 2b, Table S1). Classical methods tended to perform among the best for 3 of the 11 metrics, ML methods for 7, and LLM methods for 4. Interestingly, these metrics appeared to be performance-based for the classical methods (e.g., the number of true positives), network structure-based for the ML methods (e.g., the connectedness, number of components, graph edit distance), and both performance- and network-based for the LLM methods (e.g., number of true positive nodes, total number of edges). However, ML approaches tended to outperform LLM methods in terms of most network structure-based metrics. These

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404

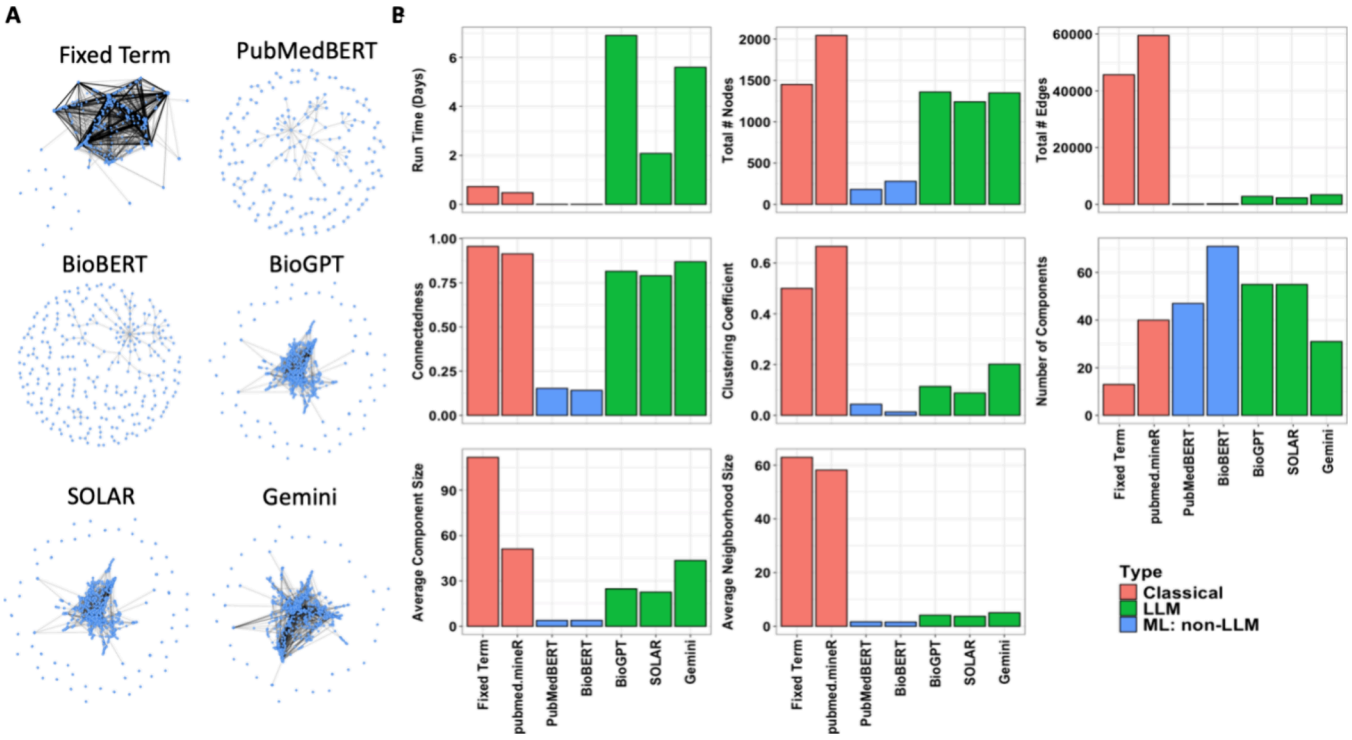


Figure 3. (A) Predicted *E. coli* protein−protein interaction networks for a selection of algorithms. (B) Metrics evaluating predicted networks with build times.

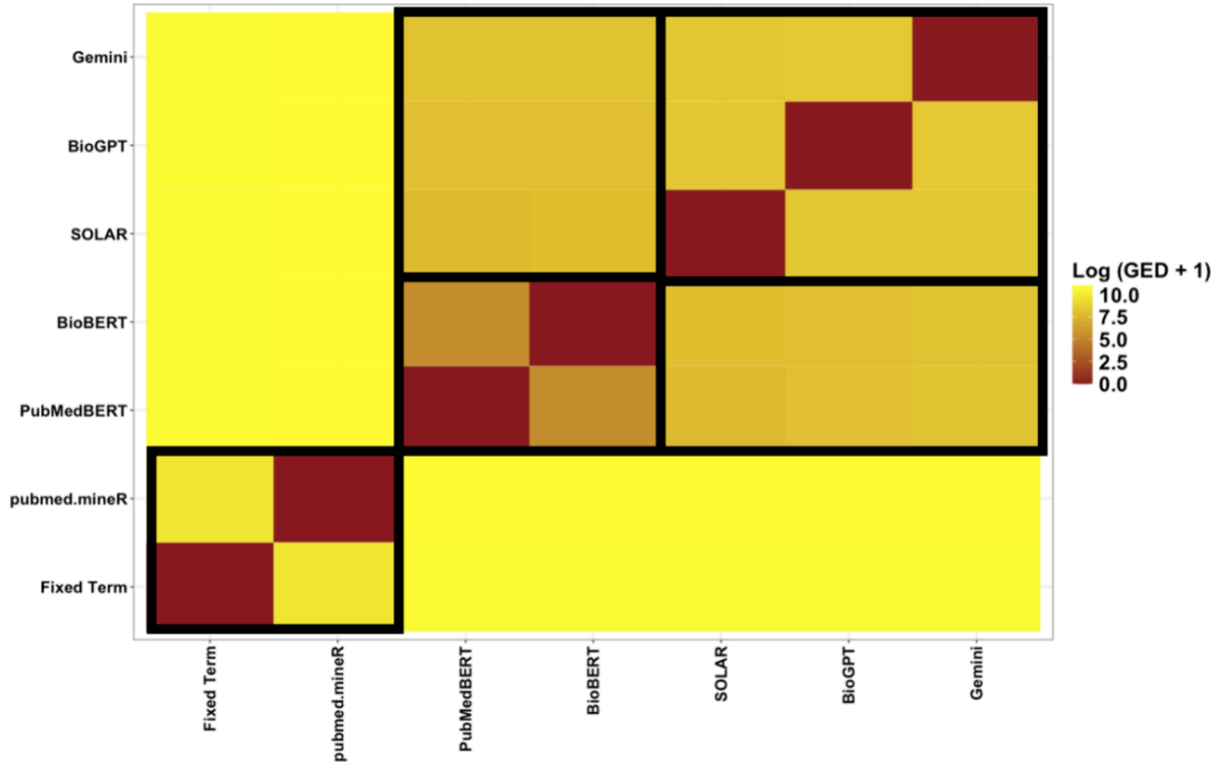


Figure 4. Log transformation of the graph edit distance (GED) of each predicted network on the *E. coli* paper cohort versus every other predicted network.

results suggest that classical approaches should be used when only detecting true positive cases is the priority, ML methods should be used when accurate network structures are the priority at the expense of true positives, and LLM methods should be used if a method between the two is desired.

In terms of varied model performance within model types, the most variability was measured within the ML approaches as opposed to the classical and LLM. The ML tools PubMedBERT, BioBERT, REACH, and TRIPS had more inconsistencies in their performances, with a delineation

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404

between the BERT models (PubMedBERT and BioBERT) and the INDRA-based models (REACH and TRIPS). BERT models were much larger (>100 nodes and >90 edges) and lightly connected with high numbers of components, while REACH and TRIPS had such small networks (<10 nodes and <5 edges) that their network properties were likely skewed to smaller values. The reduced performance in REACH and TRIPS models was likely due to their built-in biological named entity recognition, which may not be trained on *C. elegans* proteins. Another key difference was that the BERT models ran in under 5 min, while REACH took 1 h and TRIPS took 30 h. Comparing the performances of the BERT models, interestingly, BioBERT had slightly better performance than PubMedBERT with its identification of true positives and overall network structure, which contradicts the results of Lee et al.,[22] where PubMedBERT had slightly better performance than BioBERT.

Study 3: Tool Performances on the *E. coli* Paper Cohort

A selection of tools from the classical, ML, and LLM approaches were tested with 3846 publications (1160 full text and 2686 titles and abstracts) as provided by PubMed (Figure S7).

Consistent with the *C. elegans* study, the selected ML tools (both BERT models) completed in under 10 min and had the lowest number of nodes and edges with lowly connected networks and thus more network components (distinct sections of interconnected nodes). The LLM methods tended to have metrics in-between the other two tool types, and all took at least 2 days to finish their analysis, with BioGPT taking the longest at 6.9 days. The selected classical tools yielded a high number of nodes and edges in highly connected networks (Figure 3 and Table S2). Though inconsistent with the previous study, the performance of the classical methods each took over 10 h, which may be explained by their unoptimized implementations in R. Though the truth network in this case is not known, the consistent patterns demonstrate that classical methods yield high numbers of edges and nodes, ML methods yield smaller numbers of edges and nodes, and LLMs exist between the two, with more nodes than edges.

In terms of model similarity, the graph edit distance between each model network yielded two clusters—one with the classical methods and one with the ML and LLM methods (Figure 4). The BERT models had the closest GED values, forming a subcluster of the ML methods, while the LLM methods were not as close in structure to each other. GED highlights the differences and similarities between model approaches, which can be seen visually (Figure S7) and provides further validity to the decisions to present these results by their selected model types.

Across all studies, consistent patterns were observed across the tool types. Classical methods had higher rates of true positives at the expense of overconnected networks and performed best when given smaller context (e.g., titles and abstracts). ML methods had lower TPRs but overall network structures that reflected the target networks and performed best when given larger contexts (e.g., full text papers). LLM methods tended to perform between the two approaches, with TPRs and network structures between the classical and the ML approaches. Method selection therefore depends on experimental goals (e.g., desired TPR, TNR, or network structure) and contextual availability. The recommended tool per type as determined solely on the basis of this study's TPR, closest structure to the target network, and model run time was fixed term (classical method), BioBERT (ML method), and SOLAR (LLM method).

CONCLUSIONS

Here, we compared the PPI network construction capabilities of a few tools that used different predictive strategies (classical text processing, ML models, or LLMs). The selection of a specific method, therefore, depends on the needs of a particular PPI study. For example, if the goal of the PPI network is to conclusively determine protein targets, as in vaccination development, BioBERT and PubMedBERT should be used, as they offer low false positive rates. Or, if the research goal is protein function discovery, methods with higher TPRs like the classical and LLM approaches should be used. A strong caution when using LLMs, besides ensuring that there is proper hardware to run the memory-intensive algorithms, is to be wary of LLM hallucinations. Future studies could focus on the effect of utilizing more than one method based on context size (e.g., using a classical method with titles and abstracts and a ML method with full text), the feasibility of model averaging approaches, or the effect of prompt engineering and hyperparameter tuning, such as tuning the temperature parameter, to optimize LLM performances. Overall, this work has identified commonalities between tool types based on their behaviors to help researchers make informed selections on text mining tools in PPI contexts.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.4c00535.

Example of a true positive, true negative, false positive, and false negative for this study; schematic of workflows for studies 2 and 3; determining a cutoff for cosine correlation score for pubmed.mineR to separate relationships and nonrelationships; networks and metrics for the clean text versus PDF study; networks and metrics for the title and abstracts versus full text study; all networks from study 2; and all networks from study 3 (PDF)

Performance metrics of all networks in study 2 and performance metrics of all networks in study 3 (XLSX)

AUTHOR INFORMATION

Corresponding Author

Lisa M. Bramer − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0002-8384-1926; Email: Lisa.Bramer@pnnl.gov

Authors

David J. Degnan − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0001-5737-7173

Clayton W. Strauch − *AI & Data Analytics Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0002-3990-5662

Moses Y. Obiri − *Earth Systems Science Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States*

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404

Erik D. VonKaenel − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0002-8933-7413
Grace S. Kim − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States*
James D. Kershaw − *Earth Systems Science Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0009-0006-3585-8690
David L. Novelli − *AI & Data Analytics Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States*
Karl TL Pazdernik − *AI & Data Analytics Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States; Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.4c00535

Author Contributions

S.J.K, D.J.D, and M.Y.O. developed the GP-GP data set; C.W.S., J.D.K., and S.J.K. pulled PubMed publications; D.J.D. trained, implemented, and evaluated models; M.Y.O., E.D.V., D.N., K.T.P., and L.M.B. provided statistical and machine learning support; L.M.B. conceptualized and supervised the project; and D.J.D. wrote the manuscript with significant contributions from all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the PNNL Laboratory Directed Research and Development program and is a contribution of the Predictive Phenomics Initiative. PNNL is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830. The authors would like to thank the many individuals who helped identify the list of terms that indicated a biological relationship, including Sam Obermiller, Bea Meluch, Logan Lewis, Jeremy Jacobson, Natalie Winans, Anastasiya Prymolenna, Rachel Richardson, and several others.

■ ABBREVIATIONS

BA, balanced accuracy; BERT, bidirectional encoder representations from transformers;; GP-GP, gene product-gene product; GPT, generative pretrained transformer; LLM, large language model; ML, machine learning; NLP, natural language processing; PDF, portable document format; PMID, PubMed paper ID; PPI, protein−protein interaction; RE, relationship extraction; ROC, receiver operator characteristic; TNR, true negative rate; TPR, true positive rate.

■ REFERENCES

(1) Nooren, I. M.; Thornton, J. M. Diversity of protein-protein interactions. *EMBO J.* **2003**, *22* (14), 3486−3492.
(2) Zhou, Y.; Liu, Y.; Gupta, S.; Paramo, M. I.; Hou, Y.; Mao, C.; Luo, Y.; Judd, J.; Wierbowski, S.; Bertolotti, M.; Nerkar, M.; Jehi, L.; Drayman, N.; Nicolaescu, V.; Gula, H.; Tay, S.; Randall, G.; Wang, P.; Lis, J.; Feschotte, C.; Erzurum, S.; Cheng, F.; Yu, H. A comprehensive SARS-CoV-2-human protein-protein interactome reveals COVID-19 pathobiology and potential host therapeutic targets. *Nat. Biotechnol.* **2023**, *41* (1), 128−139.
(3) Safari-Alighiarloo, N.; Taghizadeh, M.; Rezaei-Tavirani, M.; Goliaei, B.; Peyvandi, A. A. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol. Hepatol. Bed. Bench.* **2014**, *7* (1), 17−31.
(4) Szklarczyk, D.; Gable, A. L.; Nastou, K. C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N. T.; Legeay, M.; Fang, T.; Bork, P.; Jensen, L.; von Mering, C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **2021**, *49* (D1), D605−D612.
(5) Del Toro, N.; Shrivastava, A.; Ragueneau, E.; Meldal, B.; Combe, C.; Barrera, E.; Perfetto, L.; How, K.; Ratan, P.; Shirodkar, G.; Lu, O.; Mészáros, B.; Watkins, X.; Pundir, S.; Licata, L.; Iannucelli, M.; Pellegrini, M.; Martin, M.; Panni, S.; Duesbury, M.; Vallet, S.; Rappsilber, J.; Ricard-Blum, S.; Cesareni, G.; Salwinski, L.; Orchard, S.; Porras, P.; Panneerselvam, K.; Hermjakob, H. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* **2022**, *50* (D1), D648−D653.
(6) UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523−D531.
(7) Chen, H.; Sharp, B. M. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinform.* **2004**, *5*, 147.
(8) Rani, J.; Shah, A. B.; Ramachandran, S. pubmed.mineR: an R package with text-mining algorithms to analyse PubMed abstracts. *J. Biosci.* **2015**, *40* (4), 671−682.
(9) Yang, Z.; Zhao, Z.; Li, Y.; Hu, Y.; Lin, H. PPIExtractor: a protein interaction extraction and visualization system for biomedical literature. *IEEE Trans. Nanobiosci.* **2013**, *12* (3), 173−181.
(10) Yang, Z.; Tang, N.; Zhang, X.; Lin, H.; Li, Y.; Yang, Z. Multiple kernel learning in protein−protein interaction extraction from biomedical literature. *AIM* **2011**, *51* (3), 163−173.
(11) Eom, J.-H.; Zhang, B.-T. Prediction of protein interaction with neural network-based feature association rule mining. In *International conference on neural information processing*; Springer: Berlin Heidelberg, 2006; pp 30−39.
(12) Xu, Q.; Hu, D. H.; Yang, Q.; Xue, H. Simpletrppi: A simple method for transferring knowledge between interaction networks for ppi prediction. *IEEE BIBMW* **2010**, 130−135.
(13) Valenzuela-Escarcega, M. A.; Babur, O.; Hahn-Powell, G.; Bell, D.; Hicks, T.; Noriega-Atala, E.; Wang, X.; Surdeanu, M.; Demir, E.; Morrison, C. T. Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database* **2018**, *2018*, bay098.
(14) Allen, J.; de Beaumont, W.; Galescu, L.; Teng, C. M. Complex event extraction using drum. *Proc. of BioNLP* **2015**, *15*, 1−11.
(15) Bachman, J. A.; Gyori, B. M.; Sorger, P. K. Automated assembly of molecular mechanisms at scale from text mining and curated databases. *Mol. Syst. Biol.* **2023**, *19* (5), No. e11325.
(16) Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM HEALTH.* **2022**, *3* (1), 1−23.
(17) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* **2020**, *36* (4), 1234−1240.
(18) Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T. Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform.* **2022**, *23* (6). DOI: .
(19) *Gemini Pro Version 1.5*; Alphabet Inc: Mountainview, CA, 2024. https://gemini.google.com/app (accessed Jun 1, 2024).
(20) Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*; Association for Computational Linguistics, 2019; pp 4171−4186.
(21) Bzdok, D.; Thieme, A.; Levkovskyy, O.; Wren, P.; Ray, T.; Reddy, S. Data science opportunities of large language models for neuroscience and biomedicine. *Neuron.* **2024**, *112* (5), 698−717.

(22) Lee, Y.; Son, J.; Song, M. BertSRC: transformer-based semantic relation classification. *BMC Med. Inform Decis Mak.* **2022**, *22* (1), 234.
(23) Krackhardt, D. Graph theoretical dimensions of informal organizations. In *Computational organization theory*; Psychology Press, 1994; pp 107−130.
(24) Butts, C. T. sna:Tools for Social Network Analysis, 2023. R package version 2.7−2. https://CRAN.R-project.org/package=sna (accessed Jan 4, 2024).
(25) Watts, D. J.; Strogatz, S. H. Collective dynamics of "small-world" networks. *Nature.* **1998**, *393* (6684), 440−442.
(26) Onnela, J. P.; Saramaki, J.; Kertesz, J.; Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **2005**, *71* (6 Pt 2), No. 065103.
(27) Fagiolo, G. Clustering in complex directed networks. *Phys. Rev. E. Stat. Nonlin. Soft Matter. Phys.* **2007**, *76* (2), No. 026107.
(28) Clemente, G. P.; Grassi, R. DirectedClustering: Directed Weighted Clustering Coefficient, 2018. R package version 0.1.1. https://CRAN.R-project.org/package=DirectedClustering (accessed Jan 4, 2024).
(29) Csardi, G.; Nepusz, T. The igraph software. *Complex Syst.* **2006**, *1695*, 1−9.
(30) Menczer, F.; Fortunato, S.; Davis, C. A. *A first course in network science*; Cambridge University Press, 2020.
(31) Sanfeliu, A.; Fu, K.-S. A distance measure between attributed relational graphs for pattern recognition. *IEEE Trans. Syst. Man. Cybern.* **1983**, *3*, 353−362.
(32) Li, Y. wrsgraph, 2024. R package version 0.0.0.9000. (accessed Apr 23, 2024).
(33) Junge, M. R.; Dettori, J. R. ROC solid: Receiver operator characteristic (ROC) curves as a foundation for better diagnostic tests. *Global Spine J.* **2018**, *8* (4), 424−429.
(34) Su, P.; Vijay-Shanker, K. Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction. *BMC Bioinform.* **2022**, *23* (1), 120.
(35) Luo, L.; Lai, P.-T.; Wei, C.-H.; Arighi, C. N.; Lu, Z. BioRED: a rich biomedical relation extraction dataset. *Brief. Bioinform.* **2022**, *23* (5), No. bbac282.
(36) Soudy, M.; Anwar, A. M.; Ahmed, E. A.; Osama, A.; Ezzeldin, S.; Mahgoub, S.; Magdeldin, S. UniprotR: Retrieving and visualizing protein sequence and functional information from Universal Protein Resource (UniProt knowledgebase). *J. Proteom.* **2020**, *213*, No. 103613.
(37) Van Rossum, G.; Drake, F. L. Python 3 Reference Manual. Create Space, 2009.
(38) PMC Open Access Submit. US National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ (accessed Apr 21, 2024).
(39) Richardson, L. beautifulsoup4, 2024. Python library version: 4.11.1. https://pypi.org/project/beautifulsoup4/ (accessed Apr 22, 2024).
(40) Ioachim, E.; Michael, M.; Stavropoulos, N. E.; Kitsiou, E.; Hastazeris, K.; Salmas, M.; Stefanaki, S.; Agnantis, N. J. Expression patterns of cyclins D1, E and cyclin-dependent kinase inhibitors p21(Waf1/Cip1) and p27(Kip1) in urothelial carcinoma: correlation with other cell-cycle-related proteins (Rb, p53, Ki-67 and PCNA) and clinicopathological features. *Urol Int.* **2004**, *73* (1), 65−73.
(41) Urade, R.; Okudo, H.; Kato, H.; Moriyama, T.; Arakaki, Y. ER-60 domains responsible for interaction with calnexin and calreticulin. *Biochem.* **2004**, *43* (27), 8858−8868.
(42) Li, P.; Jiang, H.; Yang, L.; Quan, S.; Dinocca, S.; Rodriguez, F.; Abraham, N. G.; Nasjletti, A. Angiotensin II induces carbon monoxide production in the perfused kidney: relationship to protein kinase C activation. *Am. J. Physiol. Renal. Physiol.* **2004**, *287* (5), F914−920.
(43) Moens, F.; Van den Abbeele, P.; Basit, A. W.; Dodoo, C.; Chatterjee, R.; Smith, B.; Gaisford, S. A four-strain probiotic exerts positive immunomodulatory effects by enhancing colonic butyrate production in vitro. *Int. J. Pharm.* **2019**, *555*, 1−10.
(44) Araki, K.; Hashimoto, K.; Ardyanto, T. D.; Osaki, M.; Shomori, K.; Nakamura, H.; Ito, H. Co-expression of Cox-2 and EGFR in stage I human bronchial adenocarcinomas. *Lung Cancer.* **2004**, *45* (2), 161−169.
(45) Harigai, M.; Hara, M.; Kawamoto, M.; Kawaguchi, Y.; Sugiura, T.; Tanaka, M.; Nakagawa, M.; Ichida, H.; Takagi, K.; Higami-Ohsako, S.; et al. Amplification of the synovial inflammatory response through activation of mitogen-activated protein kinases and nuclear factor kappaB using ligation of CD40 on CD14+ synovial cells from patients with rheumatoid arthritis. *Arthritis. Rheum.* **2004**, *50* (7), 2167−2177.

https://doi.org/10.1021/acs.jproteome.4c00535
J. Proteome Res. 2024, 23, 5395−5404